4^{th} International Conference on the Science of Science and Innovation, June 16-18, 2025
Copenhagen Business School, Copenhagen, Denmark

Automating Scholarly Judgment

Jiwoong Choi, Siyang Wu, Yingrong Mao, Honglin Bao

University of Chicago

Keywords: Automated Hypothesis Generation, Scientific Evaluation, Large Language Models

Background

What constitutes *good* science remains a longstanding question in both philosophy and practice. Traditional peer review, for instance, has been critiqued for subjectivity, inconsistency, and potential bias [1, 2]. Recent advances in large language models (LLMs) offer a novel opportunity to re-examine this question at scale. While past approaches often relied on citation-based metrics [4], we explore a *bottom-up* methodology that uses LLMs to generate and validate hypotheses about scientific quality—thereby shedding new light on how expert judgments might form and evolve. Here, we analyze a dataset of approximately 27K papers submitted to 45 computer science conferences, paired on review scores to create clear distinctions in perceived quality. Rather than manually defining the criteria of "good" science, we task LLMs with iteratively proposing, testing, and refining hypotheses that explain why one paper might be judged as stronger than another. Throughout this abductive reasoning process, the LLM's initial "normative" prior beliefs (e.g., a good paper has high novelty) are updated into a posterior that reflects more *professional-science* criteria (e.g., a good paper tells a good story).

Dataset and Methods

We integrated data from OpenReview and data manually scrapped from PaperCopilot.com, aligning submissions, reviewer scores, comments, and metadata, and formed pairs of papers with substantially different review scores within the conference. We then prompted an LLM to propose and test hypotheses explaining why one paper might appear stronger than the other (e.g., "Paper 1 lacks rigorous justification of its contribution").

Automated Hypotheses Generation (See Figure 1)- We begin by randomly sampling 50 paper pairs, prompting the LLM to propose 5 potential explanatory factors and testing each via repeated queries and confidence-weighted voting in [3] (i.e., "do you think this hypothesis holds true among the pair of papers? Return your judgment and the confidence about your judgment"). Any unexplained pairs become a "residual" set from which we sample another 50 pairs to generate 5 additional hypotheses, iterating until the unexplained cases are lower than 5%. This yields a final pool of 20 orthogonal hypotheses with high coverage of pairs.

Automated Hypotheses Evaluation We conducted two sets of experiments to explain these hypotheses. The first is *how the difference between two papers' embedded representations could explain the judgment*. We computed embedding differences for each paper pair and trained a Siamese neural network classifier, where two inputs share a neural network with the same parameters and weights, to detect if the judgment could be predicted by the difference in representations. We find while certain features (e.g., an elegant design) were captured well, aspects such as "weak contextualization within the broader literature" proved harder to model solely from textual embeddings. One possible reason is that these features not only require papers' own representations but also their related literature. The second set of experiments is *how much could we trust LLM's judgment*. We extract the feature from each hypothesis. We

first use LLM to annotate each paper's peer review comments regarding each feature with one of three labels: "good", "bad", and "not mention". We then ask LLM to judge the paper regarding the same set of features and retune one from three labels: "good", "bad", and "neutral". We will focus on the annotated label overlap between human judgment and LLM's judgment. Across all experiments, we use the "extended abstract" of the paper as shown in [5], which combines summarizations of the context (related literature), key ideas, methods, results, and potential future works.

Key Findings

We focus on how LLM's beliefs about "good science" are shifted during this iterative hypothesis generation and evaluation process. We collected LLM's prior, which is the set of hypotheses and their appearance frequency across 2000 generations about what is good science but without any data input. The posterior of LLM is their generated hypotheses and the proportion of pairs that each hypothesis could explain. The LLM's prior outputs often emphasized high-level, normative ideals (e.g., novelty). Through iterative refinement, however, the criteria shifted toward more professionalized norms, including *storytelling* (See Figure 2). LLMs could serve as powerful tools for uncovering latent patterns in how experts judge scientific work. Nevertheless, challenges remain. Interpretability is a critical bottleneck: while the iterative process yields human-understandable hypotheses, it relies on opaque LLM reasoning under the hood. In addition, substantial progress is still needed in guiding LLMs and humans toward a clearer understanding of what constitutes truly valuable science.

Algorithm 1: Hypothesis Search

Initialize an empty hypothesis set H and unexplained set W (all pairs);
for *each time step t* **do**
  Generate 5 hypotheses on top of the existing set H, from 50 random pairs of papers within the unexplained set W;
  Update hypothesis set H;
  Apply 5 new hypotheses to all pairs in W by 3-fold confidence-weighted voting;
  Apply 5 new hypotheses to explained pairs by 3-fold confidence-weighted voting as well;
  Update the unexplained sample set W;
  if $W < 5\%$ *of all samples* **then**
    Stop searching;
  end
end

Figure 1: Automated Hypothesis Generation Workflow

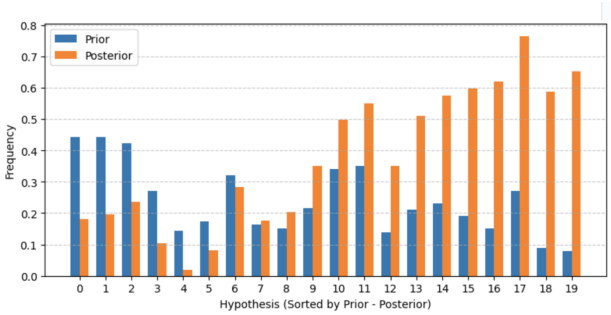

Figure 2: Prior and Posterior Distribution

References

[1] L. Bornmann. Scientific peer review. *Annual Review of Information Science and Technology*, 45(1):197–245, 2011.

[2] C. J. Lee, C. R. Sugimoto, G. Zhang, and B. Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013.

[3] A. Taubenfeld, T. Sheffer, E. Ofek, A. Feder, A. Goldstein, Z. Gekhman, and G. Yona. Confidence improves self-consistency in llms. *arXiv preprint arXiv:2502.06233*, 2025.

[4] P. Wouters et al. Rethinking impact factors. *Nature*, 569(7758):621, 2019.

[5] X. Zhang, Y. Xie, J. Huang, J. Ma, Z. Pan, Q. Liu, Z. Xiong, T. Ergen, D. Shim, H. Lee, et al. Massw: A new dataset and benchmark tasks for ai-assisted scientific workflows. *NAACL Findings*, 2025.